Nonparametric Estimation of Semiparametric Transformation Models

Jean-Pierre FLORENS
Toulouse School of Economics

Senay SOKULLU
University of Bristol

May 11, 2016

Abstract

In this paper we develop a nonparametric estimation technique for semiparametric transformation models of the form: $H(Y) = \varphi(Z) + X'\beta + U$ where H, φ are unknown functions, β is an unknown finite-dimensional parameter vector and the variables (Y, Z) are endogenous. Identification of the model and asymptotic properties of the estimator are analyzed under the mean independence assumption between the error term and the instruments. We show that the estimators are consistent, and a $\sqrt{N}$-convergence rate and asymptotic normality for $\hat{\beta}$ can be attained. The simulations demonstrate that our nonparametric estimates fit the data well.

Keywords: Nonparametric IV Regression, Inverse problems, Tikhonov Regularization, Regularization Parameter

JEL Classification: C13, C14, C30

1 Introduction

In this paper we focus on nonparametric estimation of a semiparametric transformation model. The model we study is given by the following relation:

$$H(Y) = \varphi(Z) + X'\beta + U, \quad \mathbb{E}[U|X, W] = 0 \tag{1}$$

where $Y, Z \in \mathbb{R}$ are endogenous, $X \in \mathbb{R}^q$ is exogenous, $W \in \mathbb{R}^p$ is a vector of instruments and $U \in \mathbb{R}$ is the error term. We aim to estimate the functions of interest, H and φ, and the parameter of interest, β, by nonparametric instrumental variable regression using the mean independence condition given in (1). We study the identification of the model and asymptotic properties of the estimators.

The model given in (1) is a hybrid of transformation models and partially linear models that both have been studied extensively in econometrics. Transformation models have the form $H(Y) = X'\beta + U$. These models have been used in applied econometrics not only to improve the performance of estimators but also to help to interpret the model. One well-known example is Box and Cox (1964) who propose a power transform of the dependent variable which may lead to normality in a linear regression. Horowitz (1996) gives other examples such as parametric and semiparametric proportional hazard rate models, log-linear regression and accelerated failure time models. Transformation models still get a lot of attention in econometrics, however, examples with nonparametric specifications are rare. A semiparametric partially linear model can be written as $Y = \varphi(Z) + X'\beta + U$. Use of partially linear models in applied econometrics is especially common when it is not clear how to specify the effect of one variable parametrically. Florens et al. (2012) study the estimation of β in $Y = \varphi(Z) + X'\beta + U$. Their main example is the model of Engle et al. (1986) in which the effect of temperature is specified nonparametrically in the demand for electricity. A more recent example of a partially linear specification comes from Bontemps et al. (2008) where they look at the impact of agricultural pollution on the prices of residential houses and use a nonlinear nonparametric specification for the effect of pollution.

In this paper we study the nonparametric estimation of a semiparametric transformation model in Equation (1) which includes nonparametric specifications on both sides of the equation. Hence we are extending transformation models to a general case where the transformation of the dependent variable is specified nonparametrically and the right-hand side of the equation includes a parametric and a nonparametric part. As an application of the equation we propose to estimate, we consider the estimation of demand systems in network industries, where the effect of the size of the network on demand can be ambiguous. For a brief illustration, we consider the example of the magazine market. The magazine market

is a two-sided market where the magazine is a platform serving readers and advertisers (see Kaiser and Wright, 2006; Kaiser and Song, 2009). The demands of the two end users depend on each other and hence indirect network externalities exist. Since the advertisers would like to reach as many readers as they can, they would prefer a magazine with many readers. Additionally, if the readers like advertisements, they would like to read a magazine with more advertisements. However, when the number of advertising pages increases too much in a magazine, it may have a nuisance effect on the readers and the network effect may start to decrease and even become negative. If we want to model the demand of readers for the magazine, it is better to specify this indirect network effect nonparametrically to be able to capture nonlinearities and non-monotonicities. Assuming that the demand function of readers is additive in its arguments, we can write[1]:

$$Y = F(\varphi(Z) + X'\beta + U) \tag{2}$$

where Y is the market share of the magazine on the readers' side. F is the demand function of readers. $\varphi(Z)$ is the network externality function that depends on the number of advertisements, Z, in other words, it is the effect of the number of advertisements on readers' demand for the magazine. X are observed and U are unobserved magazine characteristics for readers. In Kaiser and Song (2009), X includes number of the content pages, the cover price and the frequency of the magazine while U is assumed to be a content-related quality shock. β is a parameter to be estimated. Under the assumption that the demand function is one-to-one, we can take the inverse of it and obtain Equation (1): $F^{-1}(Y) = \varphi(Z) + X'\beta + U$ where $H(Y) = F^{-1}(Y)$. Note that this specification allows us to specify both the demand and the network effect functions nonparametrically.

To the best of our knowledge estimation of Equation (1) has not been studied, nor has it been used in an empirical application. We present the identification, estimation and asymptotic properties of this model using the mean independence condition between the error terms and instruments. So, we are not only introducing a general form of nonparametric model but also using a relatively weak condition in its estimation. The estimation we propose depends on nonparametric instrumental variable regression. It is well known in the nonparametric IV literature that this estimation problem is an ill-posed inverse problem. More broadly, the solution of our main identifying equation requires the inversion of an infinite-dimensional operator with infinitely many eigenvalues which are very close to zero. Hence, it needs a modification, or in the terminology of ill-posed inverse problems, we need to regularize the problem. In this paper, we solve the ill-posed inverse problem we encounter by *Tikhonov*

[1]For more information on the derivation of the demand equation see Bass (1969).

Regularization which can be thought of as the nonparametric counterpart of Ridge Regression. We show that we obtain consistent estimators as well as a $\sqrt{N}$-convergence rate and asymptotic normality for $\hat{\beta}$ with nonparametric IV regression.

We investigate the performance of our estimation procedure by means of a Monte Carlo simulation. Since we regularize the ill-posed inverse problem in the estimation, practical implementation requires the choice of two tuning parameters governing the bandwidth and regularization. We present a way to choose the optimal regularization parameter and use it in the simulations for a given bandwidth. Simulations show that, when the regularization parameter is chosen optimally, our estimated curves fit the actual ones well and the estimate of β is close to its true value. However, in cases where we choose the regularization parameter arbitrarily, we may have oscillating or very flat curves, as the theory suggests. This result also demonstrates the importance of the selection of the regularization parameter in ill-posed inverse problems which is encountered very often in nonparametric estimation.

This paper differs from the existing literature in the sense that it covers a general case, as it considers a semiparametric transformation model. Nonparametric IV estimation has been studied extensively; see Ai and Chen (2003), Newey and Powell (2003), Hall and Horowitz (2005), Darolles et al. (2011) and Horowitz (2011) among others. Newey and Powell (2003), Hall and Horowitz (2005), Darolles et al. (2011) and Horowitz (2011) consider models without finite-dimensional parameters and all use nonparametric IV regression to estimate the functions of interest using the mean independence condition. We also use mean independence between the error term and instruments to identify and estimate the functions and parameters of interest. Ai and Chen (2003) and Florens et al. (2012) consider the partially linear, semiparametric model. The latter uses nonparametric instrumental variables regression based on kernels and get over the ill-posed inverse problem by Tikhonov regularization, while the former restricts the set of functions to be a compact set to avoid the inverse problem and estimates the parameter and functions of interest by minimum distance sieve estimation. We follow the approach of Florens et al. (2012) and recover our functions of interest by regularizing the ill-posed inverse problem. Most of the examples of transformation models are specified and estimated parametrically. Horowitz (1996) studies semiparametric estimation of transformation models, though he makes a parametric specification for the right-hand side. Linton et al. (2008) also estimate a semiparametric transformation model but they assume a parametric transformation of the dependent variable. Chiappori et al. (2011) show the identification and estimation of a nonparametric transformation model where they specify the equations on both sides nonparametrically. In contrast to our paper, they do not have a partially linear model and they assume that the error terms are independent of the exogenous variables conditional on the endogenous variables. Feve and Florens (2010) estimate

a simplified version of our model with nonparametric instrumental regression, where they use a nonparametric transformation explained by a parametric linear model. Therefore, this paper generalizes nonparametric transformation models. Moreover, compared to the aforementioned papers, we are using a weaker assumption on the error terms and instruments.

The paper is organized as follows: In Section 2 we introduce a simple model where $X \in \mathbb{R}$ and β is normalized to 1. After studying the identification, estimation and asymptotic properties of this simpler model we generalize it to the model in Equation (1) in Section 3. A data-based method for the selection of the optimal regularization parameter is discussed in Section 4 while we present the results of a small Monte Carlo simulation exercise in Section 5. Finally, Section 6 concludes. All the proofs are presented in the appendices.

2 A Semiparametric Transformation Model and Its Nonparametric Estimation

In this section we study a simpler version of the transformation model in (1) to concentrate on the identification, estimation and asymptotic properties of the nonparametric components. In this simpler version we restrict X to be a scalar and normalize β to 1. The relationship between the variables is given by the following equation:

$$H(Y) = \varphi(Z) + X + U \tag{3}$$

$$\mathbb{E}[U|X, W] = 0$$

This is a semiparametric transformation model in which we have two endogenous variables, $Y, Z \in \mathbb{R}$, and an exogenous variable $X \in \mathbb{R}$. $U \in \mathbb{R}$ is the error term and $W \in \mathbb{R}^p$ is a vector of instruments. Here we do not need Z to be a real number although we assume so for the sake of exposition and the extension of our results to a random vector is straightforward. Moreover, although X is assumed to be a continuous random variable, the model also allows X to be discrete. Note that, in such a case, we may need further assumptions on the instruments; see Newey and Powell (2003) and Das (2005). In addition to this, X is not necessarily a scalar, and can be a vector which is indeed the case in the general model in Section 3.[2]

The variables Y, Z, X, W generate the random vector Ξ which has a cumulative distribution function F. Then for each F, we can define the subspaces of real valued functions as

[2]We thank one of the anonymous referees for pointing out these features which are allowed by our model for X and Z.

$L_F^2(Y)$, $L_F^2(Z)$, $L_F^2(X)$ and $L_F^2(W)$ which depend only on Y, Z, X and W, respectively, and which belong to a common Hilbert space denoted by L_F^2 as we assume throughout.[3] Given these, we assume that Y, Z is uniquely determined conditional on X, W. In the following, we will adopt a limited information approach by considering only part of the structural data generating process given by Equation (1).

2.1 Identification

The identification of the model is based on the conditional independence of the error term and the instruments rather than full independence. Hence, our approach differs from the existing literature not only by the nonparametric specification of functions on both sides of the transformation model but also by a weaker assumption for identification.

Consider the random vector Ξ defined above. Assume that this vector satisfies the following assumptions:

Assumption 1 *There exist two square integrable functions H and φ such that:*

$$H(Y) = \varphi(Z) + X + U$$

with

$$\mathbb{E}[U|X, W] = 0$$

We want to consider the unicity of H and φ under the mean independence condition. In order to verify this unicity we assume two regularity conditions on the joint distribution of (Y, Z, X, W).

Assumption 2 ***Completeness.*** *The distribution of (Y, Z) given (X, W) is complete in the following sense:*

$$\forall m(Y, Z) \in L_F^2(Y \times Z), \quad \mathbb{E}[m(Y, Z)|X, W] = 0 \quad a.s. \quad \Rightarrow \quad m(Y, Z) = 0 \quad a.s.$$

This assumption (also called *strong identification*) has a long history in statistics in the analysis of the relation between sufficiency and ancillarity (see Florens et al., 1990, Chapter 5) and it is essential in the study of instrumental variables estimation (see Florens et al.,

[3]In this paper, all function spaces are assumed to be L^2 spaces relative to the density of the data generating process. This choice of L^2 space is motivated by two reasons: First, the conditional expectation operator is well-defined in an L^2 space and second (different from the L^p spaces where $p \neq 2$) the L^2 spaces are Hilbert spaces which simplifies the use of adjoint operators. A theory in Banach spaces may be developed but it would be more abstract and not really motivated by applications. The choice of the density for the L^2 definition is also motivated by the simplicity of the computation of the adjoint operators.

2003; Newey and Powell, 2003; Blundell et al., 2007; Hu and Schennach, 2008; Feve and Florens, 2010; Darolles et al., 2011; Florens et al., 2012; Berry and Haile, 2014). More recently D' Haultfoeuille (2011), Hu and Shiu (2011) and Andrews (2011) have analyzed this assumption and the primitive conditions that lead to the complete distributions. In *Appendix A*, we present a discussion of preliminary conditions that lead to this assumption. As D' Haultfoeuille (2011) mentions, there is a trade-off between the regularity conditions imposed on the model and the assumptions on the nonparametric functions of interest. Contrary to control variable approaches, the completeness assumption requires no restriction on the infinite dimensional parameter of interest nor does it on the relation between the endogenous variables and the instruments. Hu and Shiu (2011) give sufficient conditions for the completeness of distributions without imposing a specific functional form. The completeness we impose is also called L^2-completeness and, as is shown by Andrews (2011), it can be obtained under mild conditions.

In addition to this, when one has enough information to make restrictions on the nonparametric function and/or the relation between the endogenous variables and the instruments, the completeness conditions can be relaxed, see D' Haultfoeuille (2011).[4] From an intuitive point of view, this assumption can be seen as the nonparametric counterpart of the rank condition in parametric IV estimation. This assumption is indeed too strong for our model because we only need the property of Assumption 2 for functions such as $m_1(Y) + m_2(Z)$. This justifies the following assumption, which is clearly weaker than the completeness condition:

Assumption 2. 1 ***Additive completeness.***

$$\forall (m_1, m_2) \in L_F^2(Y) \times L_F^2(Z), \quad E(m_1(Y)+m_2(Z)|X,W) = 0 \quad a.s. \quad m_1(Y)+m_2(Z) = 0 \quad a.s.$$

The completeness assumption is often criticized, but it should be noted that in some sense, this assumption is not necessary. We discuss this point in *Remark 7*.

Assumption 3 ***Separability.*** *Y and Z are measurably separable i.e., $\forall m(Y) \in L_F^2(Y)$ and $\forall l(Z) \in L_F^2(Z)$:*

$$m(Y) = l(Z) \Rightarrow m(.) = l(.) = constant$$

Assumption 3 is also standard in nonparametric estimation. It means that there is not an exact relation between Y and Z, or put differently, $X + U$ in Equation (3) is not equal to a constant. It is essentially a support condition on Y and Z, and it prevents the existence of a

[4]Note that it has also been shown that inference can still be done in a partially identified model when the completeness assumption does not hold; see Freyberger and Horowitz (2012) and Santos (2012).

deterministic relation between Y and Z. In particular, if the support of the joint distribution of Y and Z is the product of the two marginal supports Assumption 3 is satisfied. A more precise analysis of the measurable separability condition is given in Florens et al. (2008).

Finally, we want to normalize the function φ:

Assumption 4 ***Normalization.*** *If $\varphi(Z)$ is constant a.s. then $\varphi(Z) = 0$ a.s.*
For simplicity, we will assume that $\varphi(.)$ is normalized by the condition $\mathbb{E}[\varphi(Z)] = 0$. Under this assumption, we consider as the parameter space:

$$\mathcal{E}_0 = (H, \varphi) \in L_F^2(Y) \times L_F^2(Z) \quad such \quad that \quad \mathbb{E}[\varphi(Z)] = 0$$

Theorem 1 *Under the assumptions 1,2.1,3 and 4, the functions $H(Y)$ and $\varphi(Z)$ are identified.*

It should be noted that identification does not require the $H(Y)$ or $\varphi(Z)$ functions to be monotonic. Hence, this would allow us to discover any non-monotonicity existing in these functions. In the example of estimation of demand systems in the magazine industry, the effect of advertising pages on reader's demand might well be non-monotonic. More precisely, although readers may enjoy seeing advertisements, they may start to experience disutility if the number of advertising pages in a magazine becomes too high. An econometric specification which restricts this network effect function ($\varphi(Z)$ in equation 2) to be monotonic would not permit recovery of this non-monotonic relation, which may have important implications in terms of pricing; see Sokullu (2015). Hence, by not restricting the functions of interest to be monotonic, our identification strategy allows for recovery of these type of relations. Nonetheless, if economic theory implies that the functions of interest are monotonic, our identification assumptions can be relaxed, and identification can be achieved under much milder conditions.[5]

Furthermore, in this paper we are considering identification and estimation of a single equation. In the presence of a system of equations with non-monotonic $H(Y)$ and $\varphi(Z)$, one needs to have more assumptions to guarantee the existence of unique reduced form solutions. This has already been examined by Sokullu (2015) under a similar specification.[6]

Remark 2 *We may remark that these assumptions can be weakened in some cases. Imagine that $Y \perp W|W_1$ and $Z \perp W|W_2$ where $W = \{W_1, W_2\}$. This means that the instruments can be grouped into two components acting separately on Y and Z. We assume also that*

[5]We thank the anonymous referee for pointing this out.

[6]Indeed, in *Appendix A*, where we present an illustration of the completeness assumption, we use a system of simultaneous equations.

W_1 and W_2 are measurably separable which in particular means that W_1 and W_2 have no elements in common. In this case:

$$\mathbb{E}[H(Y) - \varphi(Z)|W] = 0 \Rightarrow \mathbb{E}[H(Y)|W_1] = \mathbb{E}[\varphi(Z)|W_2] = c$$

where c is a constant and equal to 0 because $\mathbb{E}[\varphi(z)] = 0$. Then if Y is strongly identified by W_1 and Z is strongly identified W_2, we get the identification result.

2.2 Estimation

We now continue with the estimation. Let us define the operator:

$$T : \mathcal{E}_0 = \left\{ L_F^2(Y) \times \tilde{L}_F^2(Z) \right\} \mapsto L_F^2(X, W) : T(H, \varphi) = \mathbb{E}[H(Y) - \varphi(Z)|X, W]$$

where $\tilde{L}_F^2(Z) = \{\varphi \in L_F^2(Z) | \mathbb{E}(\varphi) = 0\}$, and the inner product is defined as:

$$\langle (H_1(Y), \varphi_1(Z)), (H_2(Y), \varphi_2(Z)) \rangle = \langle H_1(Y), H_2(Y) \rangle + \langle \varphi_1(Z), \varphi_2(Z) \rangle$$

where the inner product is given by $\langle g(x), h(x) \rangle = \int_X g(x)h(x)dx$. The adjoint operator of T, T^*, satisfies:

$$\langle T(H(Y), \varphi(Z)), \psi(X, W) \rangle = \langle (H(Y), \varphi(Z)), T^*\psi(X, W) \rangle$$

for any $(H, \varphi) \in \mathcal{E}$ where $\mathcal{E} = \{L_F^2(Y) \times L_F^2(Z)\}$ and $\psi \in L_F^2(X, W)$. From this equality it follows immediately that

$$T^*\psi = (\mathbb{E}[\psi(X, W)|Y], -\mathbb{E}[\psi(X, W)|Z])$$

However, as already defined, our parameter space is $\mathcal{E}_0$. Let us denote the restriction of T to $\mathcal{E}_0$ by T_0 ($T_0 = T_{|\mathcal{E}_0}$) and the projection of $\mathcal{E}$ under $\mathcal{E}_0$ by $\mathbb{P}$, $\{\mathbb{P}(H, \varphi) = (H(Y), \varphi(Z) - \mathbb{E}[\varphi(Z)])\}$. Then the following lemma characterizes the adjoint operator T^* of T:

Lemma 3 *Let us define the operator $K : \mathcal{G} \to \mathcal{S}$ with the adjoint $K^* : \mathcal{S} \to \mathcal{G}$. Moreover, let us define $K_0 = K_{\mathcal{G}_0}$, where $\mathcal{G}_0 \in \mathcal{G}$. Then, $K_0^* = \mathbb{P}K^*$ where $\mathbb{P}$ is the projection operator that projects functions of $\mathcal{G}$ on $\mathcal{G}_0$.*

Then we can write the adjoint operator of T as:

$$T^*\psi = \begin{pmatrix} \mathbb{E}(\psi(X, W)|Y) \\ -\mathbb{P}\mathbb{E}(\psi(X, W)|Z) \end{pmatrix}$$

where $\mathbb{P}$ is the projection of $L_F^2(Z)$ on $\tilde{L}_F^2(Z)$ ($\mathbb{P}\varphi = \varphi(Z) - \mathbb{E}[\varphi(Z)]$).

The estimation problem can be written as:

$$T(H, \varphi) = X \tag{4}$$

Estimation of H and φ requires nonparametric estimation of the operator T which has an infinite dimensional range and in general is compact.[7] This, in turn, leads to an ill-posed inverse problem, since the inversion of the estimator of T leads to discontinuities of the resulting estimators with respect to the joint distribution of the data.[8] To get a stable solution, we therefore need to regularize our problem which can be done by Tikhonov Regularization. Tikhonov Regularization requires controlling the norm of the solution by a penalty term, α, which is called the *regularization parameter*.[9] The regularized solution of (4) is then given by the minimization of the following problem:

$$\min_{H,\varphi}\{\|X - T(H, \varphi)\|^2 + \alpha \|(H, \varphi)\|^2\} \tag{5}$$

Thus,

$$(H(Y), \varphi(Z))' = (\alpha I + T^*T)^{-1}T^*X \tag{6}$$

where I is the identity operator in $L_F^2(Y) \times L_F^2(Z)$. Note that the minimization is not performed on the estimated operators. Instead, we first solve the inverse problem, thus minimize the norm with a penalty and perform the estimation on the solution. The solution in (6) can be written as follows:

$$(\alpha I + T^*T)(H, \varphi) = T^*X$$

Equivalently,

$$\begin{pmatrix} \alpha H(Y) + \mathbb{E}\left[\mathbb{E}(H(Y)|X, W)|Y\right] - \mathbb{E}\left[\mathbb{E}(\varphi(Z)|X, W)|Y\right] \\ -\alpha\varphi(Z) + \mathbb{P}\mathbb{E}\left[\mathbb{E}(H(Y)|X, W)|Z\right] - \mathbb{P}\mathbb{E}\left[\mathbb{E}(\varphi(Z)|X, W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X|Y) \\ \mathbb{P}\mathbb{E}(X|Z) \end{pmatrix} \tag{7}$$

[7] The compactness is satisfied in particular when the joint density Ξ is square integrable. (See Darolles et al., 2011)

[8] Engl et al. (1996) define a problem as well-posed if the conditions below hold:
(i) For all admissible data a solution exists.
(ii) For all admissible data the solution is unique.
(iii) The solution continuously depends on the data.

[9] The choice of α is important since it characterizes the balance between the fitting and the smoothing, and in the following sections we introduce a data based selection rule for it.

The estimation of $H(Y)$ and $\varphi(Z)$ is done in two steps: First we estimate the different conditional expectations by replacing the unknown density f by a kernel estimator $\hat{f}$:

$$\hat{f}(y,z,\tilde{w}) = \frac{1}{Nh_y h_z h_{\tilde{w}}} \sum_{i=1}^{N} K\left(\frac{y-y_i}{h_y}\right) K\left(\frac{z-z_i}{h_z}\right) K\left(\frac{\tilde{w}-\tilde{w}_i}{h_{\tilde{w}}}\right)$$

where $\tilde{w}$ is the vector of instruments, K represents different kernels adapted to the dimension of the variables, h_y, h_z and $h_{\tilde{w}}$ are bandwidth parameters and N is the sample size. In the second step, we solve the system of equations given by (7) to compute the estimates $\hat{H}^\alpha(Y)$ and $\hat{\varphi}^\alpha(Z)$. Let us denote $\hat{T}$ and $\hat{T}^*$ the estimates of the operators T and T^* obtained by this plug-in method, see Feve and Florens (2010); Sokullu (2015). We can then write:

$$(\hat{H}^\alpha(Y), \hat{\varphi}^\alpha(Z)) = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^* X \tag{8}$$

Practical implementation of this method has been discussed thoroughly in Feve and Florens (2010) and Darolles et al. (2011). Especially, the implementation of the method for a semi-parametric transformation model as in this paper is discussed in detail in Sokullu (2015).[10] Moreover, similar practical implementations based on sieve approximations are studied in Horowitz (2011).

Remark 4 *Note that a single regularization parameter α is introduced in the equation system (7) for the sake of exposition. Indeed, Equation (5) can be written as:*

$$\min_{H,\varphi}\{\|X - T(H,\varphi)\|^2 + \alpha_H \|H\|^2 + \alpha_\varphi \|\varphi\|^2\}$$

which leads to a system of equations similar to (7) with α_H in the first line and α_φ in the second line. However, even if the values of the two α's are different, they should converge to zero at the same speed. In Section 4, we introduce a data based selection rule for α parameters for practical implementation.

2.3 Consistency and Rate of Convergence

In our estimation process, the functions are estimated through the estimation of the operators. For this reason, to be able to talk about the consistent estimation of the functions of interest, first we have to estimate the operators T and T^* consistently. Before stating the assumptions for consistency, let us introduce the definition of the singular value decomposition (SVD) and the operator norm we will be using throughout the paper.

[10] In contrast to this paper, the semiparametric transformation model in Sokullu (2015) has only endogenous explanatory variables.

Definition 1 *Let $\{\lambda_j, \phi_j, \psi_j\}$ be the singular system of the operator T such that:*

$$T\phi_j = \lambda_j \psi_j \quad and \quad T^*\psi_j = \lambda_j \phi_j$$

where the λ_j denote the sequence of the nonzero singular values of the compact linear operator T, ϕ_j and ψ_j, for all $j \in \mathbb{N}$, are orthonormal sequences of functions in $\mathcal{E}_0$ and $L_F^2(X, W)$, respectively. We can moreover write the singular value decomposition for each $\varphi \in \mathcal{E}_0$:[11]

$$T\varphi = \sum_{j=1}^{\infty} \lambda_j \langle \varphi, \phi_j \rangle \psi_j$$

Definition 2 *If $K : \mathcal{E}_1 \mapsto \mathcal{E}_2$ is a linear operator between the two normed spaces, then the operator norm of K is given by:*

$$\|K\| := sup\{\|K\phi\|_{\mathcal{E}_2}; \phi \in \mathcal{E}_1 \quad and \quad \|\phi\|_{\mathcal{E}_1} \leq 1\}$$

We need the following assumptions for consistency:

Assumption 5 *Source Condition: There exists $\nu > 0$ such that:*

$$\sum_{j=1}^{\infty} \frac{\langle \Phi, \phi_j \rangle^2}{\lambda_j^{2\nu}} = \sum_{j=1}^{\infty} \frac{[\langle H, \phi_{1,j} \rangle + \langle \varphi, \phi_{2,j} \rangle]^2}{\lambda_j^{2\nu}} < \infty$$

where $\Phi = (H, \varphi)$.

This assumption defines a regularity space for our functions. In other words, as stated in Carrasco et al. (2007), it can be said that the unknown value of $\Phi = (H, \varphi)$ belongs to the space Ψ_ν where

$$\Psi_\nu = \left\{ \Phi \in \mathcal{E} \quad such \quad that \quad \sum_{j=1}^{\infty} \frac{\langle \Phi, \phi_j \rangle^2}{\lambda_j^{2\nu}} < \infty \right\}$$

In fact, assuming that $\Phi \in \Psi_\nu$ just adds a smoothness condition to our functional parameter of interest. As was pointed out by Carrasco et al. (2007), this regularity assumption will give us an advantage in calculating the rate of convergence of the regularization bias. To see how this assumption works, let us consider the Fourier decomposition of Φ in the basis of ϕ_j: $\Phi = \sum_{j=1}^{\infty} \langle \Phi, \phi_j \rangle \phi_j$. For example, if ϕ_j are polynomials, the rate of decline of $\langle \Phi, \phi_j \rangle^2$ shows the accuracy of polynomial approximations of Φ. This rate has to be compared with the rate of the λ_j^2. If λ_j^2 declines exponentially fast (this case is called severely ill-posed) then Assumption 5 is strong and means that Φ is almost a polynomial. Else, the speed of

[11] For more on singular value decomposition, see Carrasco et al. (2007).

convergence will be very slow. If, on the other hand, λ_j^2 declines at a geometric rate, i.e., $\lambda_j^2 \sim \frac{1}{j^a}$, then Assumption 5 has to be interpreted as a smoothing condition, see Hall and Horowitz (2005). In particular, if $\langle \Phi, \phi_j \rangle^2 \sim \frac{1}{j^b}$, Assumption 5 is satisfied if $\nu < \frac{1}{a}(b - \frac{1}{2})$.

Assumption 6 *There exists $s \geq 2$ such that:*

- $\left\| \hat{T} - T \right\|^2 = O_p\left(\frac{1}{N h_N^{p+2}} + h_N^{2s} \right)$
- $\left\| \hat{T}^* - T^* \right\|^2 = O_p\left(\frac{1}{N h_N^{p+2}} + h_N^{2s} \right)$

where s is the minimum between the order of the kernel and the order of the differentiability of f, p is the dimension of the instrument vector W and h_N is the bandwidth.

Assumption 7

$$\left\| \hat{T}^* X - \hat{T}^* \hat{T} \Phi \right\|^2 = O_p\left(\frac{1}{N} + h_N^{2s} \right)$$

Assumptions 6 and 7 state the rates of convergence for the estimated operators. These assumptions can be satisfied once the uniform convergence of the kernel density estimator of the density of $(Y, X, W) \in \mathbb{R}^{p+1+1}$, f, is obtained. For this to hold, the preliminary conditions are needed to be imposed on the kernel functions as well as the data generating processes, see Hall and Horowitz (2005); Hansen (2008); Rothe (2010); Darolles et al. (2011). Thus, these assumptions can be shown to hold under independent and weakly dependent observations and for different types of kernel functions (see Hansen, 2008). Given the preliminary conditions, Darolles et al. (2011) show uniform convergence of f and prove the convergences of the estimated operators in Hilbert-Schmidt norm which implies convergence in the supremum norm. Intuitively, Assumption 6 means that $\|\hat{E}(H(Y)|X, W) - E(H(Y)|X, W)\|^2$ has the same behavior as the MISE of the estimate of the density of $(Y, X, W) \in \mathbb{R}^{p+1+1}$ or $\|\hat{E}(\psi(X, W)|Y) - E(\psi(X, W)|Y)\|^2$ has the same behavior as the MISE of the estimate of the density of $(Y, X, W) \in \mathbb{R}^{p+1+1}$. Let us underline that the speed of convergence in Assumption 7 is parametric $(1/N)$ up to a bias term. This is due to the fact that the nonparametric estimate of $\hat{T}\Phi$ is averaged by $\hat{T}^*$ which makes the nonparametric rate disappear. This assumption is shown to be true in regular cases in Darolles et al. (2011).

Assumption 8 $\lim_{N\to\infty} \alpha_N = 0$, $\lim_{N\to\infty} \alpha_N^2 N \to \infty$, $\lim_{N\to\infty} N h_N^{p+2} \to \infty$, $\lim_{N\to\infty} \frac{h_N^{2s}}{\alpha_N^2} = 0$, $\lim_{N\to\infty} \alpha_N^{2-\nu} N h_N^{p+2} \to \infty$ *or* $\nu \geq 2$.

Assumption 8 presents the necessary conditions for the consistency of the estimator. It is required that both α and h go to zero at some compatibility (i.e. h^{2s} should go to zero faster than α^2). The third statement in Assumption 8 is standard in kernel smoothing, and the others mean that α should not go to zero too fast relative to the estimation error.

Theorem 5 *Let us define $\Phi = (H(Y), \varphi(z))$. Let s be the minimum between the order of the kernel and the order of the differentiability of f and ν be the regularity of Φ. Under Assumptions 5 to 8:*

- $\left\|\hat{\Phi}_N^\alpha - \Phi\right\|^2 = O_p\left(\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2s}\right) + \frac{1}{\alpha^2}\left(\frac{1}{Nh_N^{p+2}} + h_N^{2s}\right)\left(\alpha^{\min\{\nu,2\}}\right) + \alpha^{\min\{\nu,2\}}\right)$
- $\left\|\hat{\Phi}_N^\alpha - \Phi\right\| \to 0$ *in probability.*

The optimal speed of convergence is obtained by the calculation of optimal α. To do this we equalize the first and the third term of the rate of convergence above, as the second term is negligible. Then we obtain the optimal α_N is proportional to $N^{-1/[\min\{\nu,2\}+2]}$. Moreover, under the assumption that $h^{2s} = O_p(1/N)$ if the conditions $[(p+2)(\nu+2)]/2\nu \leq s$ when $\nu \leq 2$, and $(p+2)/4 \leq s$ when $\nu > 2$, are satisfied, we obtain the following optimal speed of convergence:

$$\left\|\hat{\Phi}^\alpha - \Phi\right\|^2 \sim O_p\left(N^{-\frac{\min\{\nu,2\}}{\min\{\nu,2\}+2}}\right)$$

This rate of convergence follows from an argument similar to that of Darolles et al. (2011). Under more specific assumptions (for example, geometric rate of decline of λ_j, $\langle H, \phi_{1,j}\rangle$ and $\langle \varphi, \phi_{2,j}\rangle$ as in Hall and Horowitz (2005)) it may be improved upon, and shown to be minimax, see Breunig and Johannes (2009); Chen and Reiss (2011). As we want to focus on the semiparametric specification, we do not reproduce this discussion which is not specific to our model.

Remark 6 *The role of the dimension of the instruments, p, should be explained. Under the optimal choice of α, p disappears from the rate of convergence (except the fact that ν is a function of p). It has been proved in Darolles et al. (2011) that in some cases increasing the dimension of the instruments decreases the rate of decline of λ_j^2 hence increasing the ν and the rate of convergence of the estimator. The cost of increasing p is that it needs a better smoothing for the estimation of the joint density, as can be seen above from the conditions on p and s for the derivation of the optimal speed of convergence.*

Remark 7 *Let us briefly discuss our estimation and consistency result if the completeness assumption is not verified, or equivalently, if T is not one-to-one. In such a case, the null space of T, $\mathcal{N}(T)$, does not reduce to $\{0\}$. Consider its orthogonal subspace in $\mathcal{E}$, $\mathcal{N}(T)^\perp$ equal to $\overline{\mathcal{R}(T^*)}$, the closure of the range of T^*. If (H, φ) is the vector of the true values, let us denote by (H^*, φ^*), its projection in $\mathcal{E}$ on $\mathcal{N}(T)^\perp$. These two functions define the pseudo true values of the model. Using a similar proof as that of Theorem 5, it can be shown that our estimator obtained by a Tikhonov regularization is always defined even if T is not one-to-one*

and that this estimator converges to the pseudo true values under a suitable choice of the sequence of α. In other words, even if the model is not identified, the estimator converges to the best approximation of the parameter by its identified component.

3 Semiparametric Transformation Model: The General Case

This section generalizes the simple model introduced in Section 2, examines the identification and estimation of H, φ and β in Equation (1), and studies the asymptotic properties of the estimators. We show that, in semiparametric transformation models with many explanatory variables, we can obtain asymptotic normality for the estimated parameters. In other words, we show that the nonparametrically estimated parameters of a partially linear transformation model can still attain a $\sqrt{N}$-convergence rate and asymptotic normality. Remember that Equation (1) is:

$$H(Y) = \varphi(Z) + X\beta + U$$

For identification, one element of the vector β needs to be normalized to 1. Then, the model in (1) can be written as the following:

$$H(Y) = \varphi(Z) + X_0 + X_1'\beta + U \tag{9}$$

where $X = \{X_0, X_1\}$. Moreover, $Y, Z, X_0, U, V \in \mathbb{R}$, $X \in \mathbb{R}^q$ and $W \in \mathbb{R}^p$.

3.1 Identification

Identification of this general model is not very different from the previous one, nonetheless we need to make the following further assumptions:

Assumption 9 ***Conditional Additive Completeness.*** $\forall (m_1, m_2, \beta) \in L_F^2(Y) \times L_F^2(Z) \times \mathbb{R}^q \quad E(m_1(Y) + m_2(Z) + X_1'\beta | X, W) = 0 \quad a.s. \Rightarrow m_1(Y) + m_2(Z) + X_1'\beta \quad a.s.$

In this general model as well, we need the completeness assumption to identify the functions and parameters of interest. Intuitively Assumption 9 means that sets of random variables (Y, Z, X_1) and (X, W) are sufficiently correlated. Note that this assumption is implied by the fact that (Y, Z) are strongly identified by W conditional on X, i.e, $\forall g \in L_F^2(Y, Z, X_1)$, $E[g(Y, Z, X_1)|X, W] = 0$ almost surely implies $g(Y, Z, X_1) = 0$ almost surely. In *Appendix A*

we give a brief illustration of the strong identification assumption with a stronger condition on U and (X, W).

Assumption 10 *(Y, Z) and X_1 are measurably separable:*

$$m(Y, Z) = l(X_1) \Rightarrow m(.) = l(.) = constant$$

Assumption 10, *measurable separability*, requires that, in our context, there is no exact relationship between $H(Y) - \varphi(z)$ and $X_1'\beta$ which will be satisfied if $X_0 + U$ is not equal to a constant in equation (9). As already mentioned, indeed this assumption is a support condition on Y, Z and X_1. Further reference on measurable separability can be found in Florens et al. (1990).

Assumption 11 *Let Σ_{X_1} denote the variance of X_1. Then, Σ_{X_1} is positive definite.*

Theorem 8 *Under the Assumptions 1-4 and 9-11 the functions $H(Y)$ and $\varphi(Z)$ and the parameter β are identified.*

3.2 Estimation

We can now proceed with estimation. Let us keep the operator T the same as in Section 2, and introduce another operator $T_X : \mathbb{R}^{q-1} \to L_F^2(X, W) : \beta \mapsto X_1'\beta$. Equivalently its adjoint is defined $T_X^* : L_F^2(X, W) \to \mathbb{R}^{q-1} : g \mapsto \mathbb{E}[X_1 g(X, W)]$ which follows from the following relation:

$$\langle T_X\beta, g(X, W)\rangle = \langle \beta, T_X^* g(X, W)\rangle$$

Then we can write:

$$T(H, \varphi) - T_X\beta = X_0 \tag{10}$$

The normal equations are:

$$T^*T(H, \varphi) - T^*T_X\beta = T^*X_0 \tag{11}$$

$$T_X^*T(H, \varphi) - T_X^*T_X\beta = T_X^*X_0 \tag{12}$$

From Equation (12), we get $\beta = (T_X^*T_X)^{-1}T_X^*T(H, \varphi) - (T_X^*T_X)^{-1}T_X^*X_0$. If we substitute it into Equation (11), we obtain an expression for (H, φ) in the general case:

$$(H(Y), \varphi(Z)) = (\alpha I + T^*(I - P_X)T)^{-1}T^*(I - P_X)X_0$$

where $P_X = T_X(T_X^*T_X)^{-1}T_X^*$. Equivalently[12]:

$$\begin{pmatrix} \alpha H + \mathbb{E}\left[(I-P_X)\mathbb{E}(H|X,W)|Y\right] - \mathbb{E}\left[(I-P_X)\mathbb{E}(\varphi|X,W)|Y\right] \\ -\alpha\varphi + \mathbb{PE}\left[(I-P_X)\mathbb{E}(H|X,W)|Z\right] - \mathbb{PE}\left[(I-P_X)\mathbb{E}(\varphi|X,W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[(I-P_X)X|Y\right] \\ \mathbb{PE}\left[(I-P_X)X|Z\right] \end{pmatrix} \tag{13}$$

As explained in Section 2, conditional expectations can be replaced by kernel estimators to get the estimates of H and φ in the second step.[13] Once $(\hat{H}^\alpha(Y), \hat{\varphi}^\alpha(Z))$ is obtained, it can be replaced back in Equation (12) to get an estimate of β.

$$\hat{\beta} = (\hat{T}_X^*\hat{T}_X)^{-1}\hat{T}_X^*(\hat{T}(\hat{H}^\alpha, \hat{\varphi}^\alpha) - X_0)$$

Note that we can also directly estimate β by substituting (H, φ) from Equation (11) into Equation (12). Then this will lead to:

$$\hat{\beta} = \left(\hat{T}_X^*\hat{T}(\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}_X - \hat{T}_X^*\hat{T}_X\right)^{-1}\left(\hat{T}_X^* - \hat{T}_X^*\hat{T}(\alpha_N I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\right)X_0 \tag{14}$$

3.3 Asymptotic Properties of $\hat{\beta}$

We now continue with the asymptotic properties of $\hat{\beta}$. In a semiparametric context the $\sqrt{N}$-convergence of the parametric component is a standard question (see Ichimura and Todd, 2007), and is generally addressed in cases where the nonparametric component is a density or a regression function. Usually this $\sqrt{N}$-convergence requires assumptions to distinguish the nonparametric and parametric part of the model. In this paper, the nonparametric component is estimated by solving an inverse problem which complicates obtaining the $\sqrt{N}$-convergence rate for $\hat{\beta}$ even further. In the sequel we present the assumptions that are needed to prove the parametric rate of convergence and asymptotic normality. Note that once we show that we can obtain $\sqrt{N}$-consistency for $\hat{\beta}$, the consistency of $(\hat{H}, \hat{\varphi})$ follows from Section 2 in a straightforward way.

Let $\{\lambda_j, \phi_j, \psi_j\}$ for $j \geq 1$ be the singular system of the operator T as defined before and let $\{\mu_l, e_l, \tilde{\psi}_l\}$ for $l = 1, 2, .., q-1$ be the singular system of the operator T_X, such that for each $\beta \in \mathbb{R}^{q-1}$ we can write:

$$T_X\beta = \sum_{l=1}^{q-1} \mu_l \langle \beta, e_l \rangle \tilde{\psi}_l$$

[12] In Equation (13) we denote $H(Y)$ by H and $\varphi(Z)$ by φ for the sake of exposition.

[13] Note that the α in this generalized version and the α in the simple version need not necessarily be the same.

Assumption 12 *Source Condition: There exists $\eta > 0$ such that:*

$$\max_{l=1,\dots,q-1} \sum_{j=1}^{\infty} \frac{\left\langle \tilde{\psi}_l, \psi_j \right\rangle^2}{\lambda_j^{2\eta}} < \infty$$

This source condition explains the collinearity between (Y, Z) and (X_1). Indeed, Assumption 12 is false if the range of T is included in the linear space generated by the elements of X_1. In contrast, if the range of T (the space of $\mathbb{E}(H(Y)|X, W) - \mathbb{E}(\varphi(Z)|X, W)$ for all H and φ) is orthogonal to X_1, then Assumption 12 is directly satisfied because the term $\left\langle \tilde{\psi}_l, \psi_j \right\rangle$ cancels out. This assumption says that the degree of collinearity is not too high compared to the singular values of T. In other words, any linear function of X_1 has Fourier coefficients in the basis ψ_j declining sufficiently fast. In fact, the values of η give a measurement of collinearity, i.e., $\eta = 0$ if there is perfect collinearity and $\eta = \infty$ if there is no collinearity. Assumption 12 is important because it guarantees that the nonparametric rate of the estimation of $H(Y)$ and $\varphi(Z)$ does not contaminate the parametric rate of $\hat{\beta}$.

Assumption 13 *Parameters given in the Source Conditions in Assumptions 5 and 12 are both greater than or equal to two, i.e., $\nu \geq 2$ and $\eta \geq 2$.*

We make this assumption for the sake of exposition of the theorem and the proof, without loss of generality. Remember that we use Tikhonov Regularization to regularize the ill-posed inverse problem we encounter. Since the Tikhonov Regularization has a qualification of two, we can not improve upon the rate of convergence when the functions we consider have regularity greater than 2, i.e., $\nu, \eta > 2$. As a result, regularization bias in Section 2 ($\|(\alpha I + T^*T)^{-1}T^*T\Phi - \Phi\|^2$) is of order $O(\alpha^{\min\{\nu,2\}})$, which would be $O(\alpha^2)$ if $\nu \geq 2$. So, by introducing Assumption 13, we get rid of the min notation.

Assumption 14 $\lim_{N\to\infty} N\alpha \to 0$, $\lim_{N\to\infty} N\alpha_N h_N^{2s} \to 0$, $\lim_{N\to\infty} \frac{\alpha_N}{h_N^{p+q+1}} \to 0$.

Assumption 14 presents the conditions to obtain asymptotic normality of the $\hat{\beta}$.

We also modify Assumptions 6 and 8 to account for the change in the dimension of X.

Assumption 6. 1 *There exists $s \geq 2$ such that:*

- $\left\|\hat{T} - T\right\|^2 = O_p\left(\frac{1}{Nh_N^{p+q+1}} + h_N^{2s}\right)$
- $\left\|\hat{T}^* - T^*\right\|^2 = O_p\left(\frac{1}{Nh_N^{p+q+1}} + h_N^{2s}\right)$

where s is the minimum between the order of the kernel and the order of the differentiability of f, p is the dimension of the instrument vector W, q is the dimension of X and h_N is the bandwidth.

Assumption 8. 1 $\lim_{N\to\infty}\alpha_N \to 0$, $\lim_{N\to\infty} h_N^{2s} \to 0$, $\lim_{N\to\infty} N h_N^{p+q+1} \to \infty$.

Let us denote $\mathcal{R}(T)$ the range of T and $\mathcal{R}(T)^\perp$ its orthogonal space in $L_F^2(X,W)$. The null space of T^* is denoted by $\mathcal{N}(T^*)$. We assume that the set of instruments is sufficiently rich such that:

Assumption 15 $\mathcal{R}(T)^\perp = \mathcal{N}(T^*) \neq \{0\}$.

In practice, this assumption implies that there exists an element ψ_j defined by the SVD of T such that $\psi_j \in \mathcal{R}(T)^\perp$. For example, this condition is satisfied in the joint nondegenerate normal case, i.e, if (Y,Z,X,W) is jointly distributed as a nondegenerate normal distribution. In such a case, the null space of T^* is $\{0\}$ if the range of the covariance with (Y,Z) and (X,W) is equal to the dimension of (X,W). Note that this is impossible even if $X_0, X_1 \in \mathbb{R}$ and W has at least one element.

Assumption 16 *For $\delta > 0$, we have:*

- $\mathbb{E}[|U|^{2+\delta}\,|X,W] = c$, *for any* $c \in \mathbb{R}$
- $\mathbb{E}[|(I - P_{YZ})X_1|^{2+\delta}] < \infty$ *where* $P_{YZ} = T(T^*T)^{-1}T^*$

Assumption 16 gives the conditions needed to satisfy the Liapounoff condition to apply the Liapounoff central limit theorem to obtain the asymptotic normality.

Using equation (14) we can show that:
$\sqrt{N}(\hat\beta - \beta) = \hat M_\alpha^{-1}\left\{\sqrt{N}[T_X^*(I - P_{YZ})\hat E(U|X,W)] + O_p(1)\right\}$ where $\hat M^\alpha = \hat T_X^*\hat T(\alpha I + \hat T^*\hat T)^{-1}\hat T^*\hat T_X - \hat T_X^*\hat T_X$, $P_{YZ} = T(T^*T)^{-1}T^*$, and $\hat E(U|X,W) = X_0 - \hat T(H,\varphi) + \hat T_X\beta$.[14] Given this decomposition and the assumption that $Var[U|X,W] = \sigma^2$, the asymptotic variance of $\hat\beta$ will be given by $N^{-1}\sigma^2 M^{-1}[\sum_{j/\psi_j\in\mathcal{R}(T)^\perp} E(X_1\psi_j)E(X_1\psi_j)']M^{-1}$ where $M = T_X^*T(T^*T)^{-1}T^*T_X - T_X^*T_X$ and $\|\hat M_\alpha^{-1} - M^{-1}\| = o_p(1)$. The next theorem formalises this:

Theorem 9 *Assume that $Var[U|X,W] = \sigma^2$. Moreover assume that Assumptions 5, 6.1, 7, 8.1, 12, 13, 14, 15 and 16 hold. Then:*

$$\sqrt{N}(\hat\beta - \beta) \to \mathcal{N}(0, V)$$

where $V = \sigma^2 M^{-1}[\sum_{j/\psi_j\in\mathcal{R}(T)^\perp} E(X_1\psi_j)E(X_1\psi_j)']M^{-1}$ and $M = T_X^*T(T^*T)^{-1}T^*T_X - T_X^*T_X$ and $\psi_j \in \mathcal{R}(T)^\perp$.

Theorem 9 shows that a $\sqrt{N}$-convergence rate and asymptotic normality for $\hat\beta$ can be obtained. Note that if the range of operators T and T_X are orthogonal to each other, the terms T^*T_X and T_X^*T in normal equations (11) and (12) will vanish and the estimation of β will not be affected by the presence of the nonparametric part.

[14] See *Appendix B.*

4 Data Based Selection of α_N

The regularization parameter plays a crucial role in estimation since it balances fitting and smoothing. For this reason, the choice of regularization parameter is very important in practice. In the case of an arbitrary choice, a regularization parameter which is too high will lead to very flat estimated curves whereas a regularization parameter which is too small results in highly oscillating estimated curves.

Morozov (1993) and Engl et al. (1996) propose a heuristic selection rule for α, called the discrepancy principle. The discrepancy principle is based on the comparison between the residual of the functional equation and the assumed bound for the noise level. It has been proven that the regularization method where α is defined via this rule is convergent and of optimal order. Following the discrepancy rule, Feve and Florens (2010) and Darolles et al. (2011) suggest a data driven selection method for α in nonparametric instrumental variables estimation. In both papers, the idea depends on the minimization of a function of the squared norm of residuals. The squared norm of residuals can be shown to reach its minimum at $\alpha_N = 0$, hence it cannot be used directly and a function of it is needed. This function is obtained first by calculating the residuals from an estimation obtained by *Tikhonov Regularization of order 2* and second by dividing the squared norm of the residuals either by α_N^2 as in Darolles et al. (2011) or by α_N as in Feve and Florens (2010). One of the drawbacks of Tikhonov regularization (of order one) is that its qualification is two, when the function being estimated is very regular, i.e. $\nu > 2$, one cannot improve more on the rate of convergence. Obtaining the residuals from a second order Tikhonov Regularization is especially done for cases where $\nu > 2$.[15] The practical implementation of this selection rule is done as follows: Given a grid of α's, i.e., $\alpha \in \mathcal{A}$, the value of the function of the squared residuals is computed for each α on the grid. Then, the regularization parameter which gives the minimum value for that function is picked as the optimal one:

$$\alpha_N^* = \underset{\alpha \in \mathcal{A}}{\operatorname{argmin}} \frac{1}{\alpha^2} \|\hat{\epsilon}_{(2)}^{\alpha}\|^2$$

where $\hat{\epsilon}_{(2)}^{\alpha}$ is the vector of residuals of the estimation regularized by Tikhonov Regularization of order 2.

The extension of this proposed data-driven selection rule of α to our case is not straightforward as we have to pick two regularization parameters to estimate our semiparametric transformation model, see Equations (7) and (13). These two regularization parameters for two unknown functions need not necessarily be the same. To get over this problem we first

[15] For more information on Iterated Tikhonov Regularization see Engl et al. (1996), Chapter 5.

assume that there is a constant ratio between two regularization parameters, i.e. $\alpha_\varphi = c\alpha_H$ for $c > 0$.[16] We then propose to choose optimal values for α_H and c in two steps. Below we explain these steps for both the simple model and the general model.

- **Simple Model**

  Let us re-write the model given in equation (3) as follows: $G(Y,Z) = X + U$ where $G(Y,Z) = H(Y) - \varphi(Z)$. In the first step we pick α_H using the data based selection rule defined in Darolles et al. (2011) as if we are estimating the function G. Under the mean independence condition in (1) the main identifying equation can be written as $T_G G(Y,Z) = X$ where $T_G : L_F^2(Y,Z) \mapsto L_F^2(X,W) : T_G G = \mathbb{E}[G(Y,Z)|X,W]$. The adjoint T_G^* is defined as: $T_G^* : L_F^2(X,W) \mapsto L_F^2(Y,Z) : T_G^* \phi = \mathbb{E}[\phi(X,W)|Y,Z]$. Then $\hat{G}$ is given by:

$$\hat{G}^\alpha(Y,Z) = (\alpha I + \hat{T}_G^* \hat{T}_G)^{-1} \hat{T}_G^* X$$

  and $\hat{G}$ obtained from estimation with Tikhonov regularization of order 2 is given by:

$$\hat{G}_{(2)}^\alpha(Y,Z) = (\alpha I + \hat{T}_G^* \hat{T}_G)^{-1} (\hat{T}_G^* X + \alpha \hat{G}_{(1)}^\alpha)$$

  The residuals from the estimation obtained by Tikhonov Regularization of order 2 are computed as:

$$\hat{\epsilon}_{(2)}^\alpha = \hat{T}_G^* X - \hat{T}_G^* \hat{T}_G \hat{G}_{(2)}^\alpha$$

  Then the optimal α_H is given by the minimization of the following problem:

$$\alpha_H^* = \underset{\alpha}{\operatorname{argmin}} \frac{1}{\alpha^2} \|\hat{\epsilon}_{(2)}^\alpha\|^2$$

  In the second step, we plug α_H^* in our original estimation problem:

$$\begin{pmatrix} \alpha_H^* H(Y) + \mathbb{E}\left[\mathbb{E}(H(Y)|X,W)|Y\right] - \mathbb{E}\left[\mathbb{E}(\varphi(Z)|X,W)|Y\right] \\ -\alpha_H^* c\varphi(Z) + \mathbb{PE}\left[\mathbb{E}(H(Y)|X,W)|Z\right] - \mathbb{PE}\left[\mathbb{E}(\varphi(Z)|X,W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X|Y) \\ \mathbb{PE}(X|Z) \end{pmatrix} \tag{15}$$

  and choose c to minimize the squared norm of residuals obtained from an estimation regularized by Tikhonov regularization of order 2.

- **General Model**

  In the general case as well, we first choose α_H as if we are estimating a function of (Y,Z) instead of two separate functions $H(Y)$ and $\varphi(Z)$. We then replace it in

[16] α_φ represents the α in front of φ and α_H represents the α in front of H in Equation (7).

the original estimation equation and optimize over c. Let us re-write the model as: $G(Y,Z) = X_0 + X_1'\beta + U$. Using the mean independence condition given in (1) and the operators T_G and T_X which are already defined, we can write: $T_G G(Y,Z) = X_0 + T_X\beta$ which will lead to the following normal equations:

$$T_G^* T_G G(Y,Z) = T_G^* X_0 + T_G^* T_X \beta \tag{16}$$

$$T_X^* T_G G(Y,Z) = T_X^* X_0 + T_X^* T_X \beta \tag{17}$$

We use Equation (17) to get an expression for β and replace it in Equation (16). Then $\hat{G}^{\alpha}_{(1)}$ is given by:

$$\hat{G}^{\alpha}_{(1)}(Y,Z) = (\alpha I + \hat{T}_G^*(I - \hat{P}_X)\hat{T}_G)^{-1}\hat{T}_G^*(I - \hat{P}_X)X_0$$

where $\hat{P}_X = \hat{T}_X(\hat{T}_X^*\hat{T}_X)^{-1}\hat{T}_X^*$. The Tikhonov regularized estimator of order 2 can be written as:

$$\hat{G}^{\alpha}_{(2)}(Y,Z) = (\alpha I + \hat{T}_G^*(I - \hat{P}_X)\hat{T}_G)^{-1}(\hat{T}_G^*(I - \hat{P}_X)X_0 + \alpha\hat{G}^{\alpha}_{(1)})$$

We can then write the residuals as:

$$\hat{u}^{\alpha}_{(2)} = \hat{T}_G^*(I - \hat{P}_X)X_0 - (\hat{T}_G^*(I - \hat{P}_X)\hat{T}_G)\hat{G}^{\alpha}_{(2)}$$

The optimal α_H in the general setting is the argument that minimizes the following:

$$\alpha_H^* = \underset{\alpha}{\operatorname{argmin}} \frac{1}{\alpha^2}\|\hat{u}^{\alpha}_{(2)}\|^2$$

The second step consists of replacing the α_H^* in the original problem below and minimizing the squared norm of residuals by choosing the optimal c.[17]

$$\begin{pmatrix} \alpha_H^* H + \mathbb{E}\left[(I - P_X)\mathbb{E}(H|X,W)|Y\right] - \mathbb{E}\left[(I - P_X)\mathbb{E}(\varphi|X,W)|Y\right] \\ -\alpha_H^* c\varphi + \mathbb{PE}\left[(I - P_X)\mathbb{E}(H|X,W)|Z\right] - \mathbb{PE}\left[(I - P_X)\mathbb{E}(\varphi|X,W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[(I - P_X)X|Y\right] \\ \mathbb{PE}\left[(I - P_X)X|Z\right] \end{pmatrix} \tag{18}$$

One last issue in the choice of regularization parameter is its sensitivity to the choice of bandwidth. The method of residuals is defined for given bandwidths. In our simulations, we choose the bandwidth by a rule of thumb and then optimize on the regularization parameter.

[17] In Equation (18) we denote $H(Y)$ by H and $\varphi(Z)$ by φ for the sake of exposition.

Feve and Florens (2010) use an iterative approach in their simulations where they first choose α for an arbitrary bandwidth and then they iterate the optimization to choose the bandwidth. They conclude that the results do not change drastically when an iterative scheme is used since α adapts itself for any a priori selection of bandwidth. The simultaneous choice of the regularization parameter and bandwidth is still an open question in the literature and is left for future work.

5 A Simulation Analysis

This section presents a Monte Carlo simulation analysis of our estimation method. We first explain the data generating process and then present our results.

We simulate the following model:

$$H(Y) = \varphi(Z) + X_0 + X_1\beta + U \tag{19}$$

$H(Y)$ is chosen to be the inverse of the logistic survival function, i.e., $H(Y) = S^{-1}(Y)$ where $S^{-1}(Y) = log((1-Y)/kY)$. Moreover $\varphi(Z)$ is chosen to be: $\varphi(Z) = Z^2 + b$, where $b = -\mathbb{E}(Z^2)$. This gives us a φ function with zero mean which satisfies Assumption 4. Then the simulated semiparametric transformation model is given by the following:

$$log\left(\frac{1-Y}{kY}\right) = Z^2 + b + X_0 + X_1'\beta + U \tag{20}$$

We associate the parameters with the following values: $k = 1$ and $\beta = 0.3$. X_0, X_1, Z and the instrument W are real numbers. X_0 and X_1 are exogenous and are drawn independently from a standard uniform distribution and so is W. We then draw U from a normal distribution with mean 0 and variance $(X_0 + X_1 + W)/70$. This variance leads to a model where $Var(U)/Var(H(Y)) = 0.17$. Z is constructed as $Z = 0.2W + \eta_W$ where η_W is generated such that Z is endogenous: $\eta_W = 0.5U + \epsilon_W$, with $\epsilon_W \sim \mathcal{N}(0, 0.04)$.

We perform the simulation for different sample sizes to control for the effect of sample size on the estimator properties. We generate 1000 samples of sizes 200, 500 and 1000.

In the estimation process, all the kernels are Gaussian and the bandwidths of the kernels are computed by a rule of thumb. For the regularization parameter, we use the data-based selection rule defined in Section 5 for each estimation. The simulation is performed by author written code in MATLAB.

Results are given in Table 1 and Figures C1 to C4 in *Appendix C*. Table 1 shows the results for the estimate of β for different sample sizes. We report the means and the standard

errors of the estimator. As can be seen from the table, the nonparametric IV estimates of β are not far from its true value, 0.3. Figure C1 shows the histogram of $\hat{\beta}$ obtained from 1000 Monte Carlo simulations of a sample of 1000 observations. The histogram looks quite similar to the pdf of a normal distribution which complements the asymptotic normality result in Section 3.3 empirically.

Figures C2 to C4 show the estimates of nonparametric functions in our model. Figure C2 shows the estimated functions over the true ones for a single sample of 500 observations whereas Figure C3 presents a Monte Carlo analysis for a sample size of 500. Our results are satisfactory. Both figures show that we can estimate the nonparametric function well and get close to the true functions. Finally, Figure C4 shows the estimates for a sample size of 500 for an arbitrary selection of regularization parameters, and indicates the importance of using our data driven selection rule.

6 Conclusion

In this paper, we have considered the nonparametric estimation of a semiparametric transformation model. The equation we introduce is motivated by the empirical study of network industries but can be applied to many economic models. We have studied the identification and estimation of the model, the asymptotic properties of the estimators, and presented a data-based selection rule for the regularization parameter for a fixed bandwidth. Development of a rule for the simultaneous selection of the two tuning parameters is left for future work.

The contributions of the paper are many-fold. First, it considers a transformation model where both left-hand side and right-hand side functions are introduced nonparametrically. Second, for the right-hand side, we adopt a partially linear specification and show that we can obtain asymptotic normality in the nonparametric estimation of the parametric part. Third, all the results of this general model hold under the assumption of mean independence which is weaker than a full independence condition.

Other extensions are possible. First of all, estimation of a system of semiparametric transformation models with a full information approach is worth studying. Moreover, the estimation method and its asymptotic properties can be generalized to nonparametric techniques other than kernels, which would be useful when working with high dimensional variables. Nonparametric tests of specification for transformation models are still underdeveloped in the literature. Finally, estimation of a structural economic model by applying the method developed here would be an interesting application.

References

Ai, C. and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Amemiya, T. (1986): *Advanced Econometrics*, Oxford: Basil Blackwell.

Andrews, D. W. (2011): "Examples of L^2-Complete and Boundedly-Complete Distributions," Cowles Foundation Discussion Papers 1801, Cowles Foundation for Research in Economics, Yale University.

Bass, F. M. (1969): "A New Product Growth for Model Consumer Durables," *Management Science*, 15, 215–227.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

Bontemps, C., M. Simioni, and Y. Surry (2008): "Semiparametric hedonic price models: assessing the effects of agricultural nonpoint source pollution," *Journal of Applied Econometrics*, 23, 825–842.

Box, G. E. P. and D. R. Cox (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society*, 26, 211–252.

Breunig, C. and J. Johannes (2009): "On rate optimal local estimation in nonparametric instrumental regression," Tech. rep., Heidelberg University.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 6 of *Handbook of Econometrics*, chap. 77.

Chen, X. and M. Reiss (2011): "On Rate Optimality For Ill-Posed Inverse Problems In Econometrics," *Econometric Theory*, 27, 497–521.

Chiappori, P.-A., I. Komunjer, and D. Kristensen (2011): "Nonparametric Identification and Estimation of Transformation Models," CAM Working Papers 2011-01, University of Copenhagen. Department of Economics. Centre for Applied Microeconometrics.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565.

Das, M. (2005): "Instrumental variables estimators of nonparametric models with discrete endogenous regressors," *Journal of Econometrics*, 124, 335–361.

D' Haultfoeuille, X. (2011): "On The Completeness Condition In Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471.

Engl, H. W., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*, Dordrecht: Kluwer Academic Publications.

Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986): "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–20.

Feve, F. and J.-P. Florens (2010): "The Practice of Non Parametric Estimation by Solving Inverse Problems: The Example of Transformation Models," *Econometrics Journal*, 13.

Florens, J.-P., J. Heckman, C. Meghir, and E. Vytlacil (2003): "Instrumental Variables, Local Instrumental Variables and Control Functions," Tech. rep.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206.

Florens, J.-P., J. Johannes, and S. V. Bellegem (2012): "Instrumental regression in partially linear models," *Econometrics Journal*, 15, 304–324.

Florens, J.-P., M. Mouchart, and J.-M. Rolin (1990): *Elements of Bayesian Statistics*, New York: M. Dekker.

Freyberger, J. and J. Horowitz (2012): "Identification and Shape Restrictions in Nonparametric Insturmental Variables Estimation," Tech. rep.

Hall, P. and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 32, 2904–2929.

Hansen, B. E. (2008): "Uniform Convergence Rates For Kernel Estimation With Dependent Data," *Econometric Theory*, 24, 726–748.

Horowitz, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–137.

——— (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79, 347–397.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y. and J.-L. Shiu (2011): "Nonparametric identification using instrumental variables: sufficient conditions for completeness," CeMMAP working papers CWP25/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Ichimura, H. and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. Heckman and E. Leamer, Elsevier, vol. 6 of *Handbook of Econometrics*, chap. 74.

Kaiser, U. and M. Song (2009): "Do media consumers really dislike advertising? An empirical assessment of the role of advertising in print media markets," *International Journal of Industrial Organization*, 27, 292–301.

Kaiser, U. and J. Wright (2006): "Price structure in two-sided markets: Evidence from the magazine industry," *International Journal of Industrial Organization*, 24, 1–28.

Linton, O., S. Sperlich, and I. V. Keilegom (2008): "Estimation of a Semiparametric Transformation Model," *Annals of Statistics*, 36, 686–718.

Morozov, V. A. (1993): *Regularization Methods for Ill-Posed Problems*, Florida: CRC Press.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Pollard, D. (1984): *Convergence of Stochastic Processes*, New York: Springer-Verlag.

Rothe, C. (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, 155, 56–70.

Santos, A. (2012): "Inference in Nonparametric Instrumental Variables With Partial Identification," *Econometrica*, 80, 213–275.

Sokullu, S. (2015): "A Semi-parametric Analysis of Two-Sided Markets: An Application to the Local Daily Newspapers in the U.S." *Journal of Applied Econometrics*, to appear.

Appendices

A Illustration of the Completeness Assumption

Assumption 2 (Assumptions 2.1 and 9 as well)or the so-called *Completeness* assumption is crucial in the identification of nonparametric IV models. In this section, we will give an illustration of primitive conditions needed in the case of a simultaneous equations system as in the model in Sokullu (2015). Let us consider the model:

$$H(Y) = \varphi(Z) + X_1\beta + X_0 + U \tag{21}$$

$$G(Y, Z, X_1) = W + V \tag{22}$$

where $Y, Z, X_1, X_0, W, U, V \in \mathbb{R}$. Let us define the following variables:

$$\zeta = H(Y) - \varphi(Z) - X_1\beta$$

$$\eta = G(Y, Z, X_1)$$

To show that (Y, Z, X_1) is complete for (X, W) where $X = (X_0, X_1)$ we need the following assumptions:

Assumption 17 *The function $h : (Y, Z, X_1) \mapsto (\zeta, \eta, X_1)$ is a bijection.*

Assumption 18 *(U, V) is independent of (X, W).*

Assumption 19 *The Fourier transform of joint distribution of (U,V) is different from 0, i.e.,*

$$\bar{\mathcal{F}}_f(t, s) = \int\int e^{-i(t\mu+s\nu)} f_{u,v}(\mu, \nu) d\mu d\nu \neq 0$$

where $\bar{\mathcal{F}}_f$ is the complex conjugate of $\mathcal{F}_f$.

Lemma 10 *Under assumptions 17-19, (Y, Z) is complete for (X, W).*

Proof. Under Assumption 17, if (ζ, η, X_1) is complete for (X, W), then (Y, Z, X_1) is complete for (X, W) as well (see Florens et al., 1990, Chapter 5). Hence, it is enough to show that (ζ, η, X_1) is complete for (X, W), i.e:

$$If \quad \mathbb{E}[\phi(\zeta, \eta, X_1)|X, W] = 0 \quad a.s. \Rightarrow \phi(\zeta, \eta, X_1) = 0 \quad a.s$$

Let us write the expectation:

$$\int\int \phi(\zeta, \eta, X_1) f_{\zeta,\eta}(\zeta, \eta|X, W) d\zeta d\eta = 0$$

By Assumption 18:

$$\forall X_0, X_1, W \quad \int\int \phi(\zeta, \eta, X_1) f_{U,V}(\zeta - X_0, \eta - W) d\zeta d\eta = 0$$

We apply the Fourier Transform:

$$\int\int\int\int\int e^{i(tX_0+sW+rX_1)} \phi(\zeta, \eta, X_1) f_{U,V}(\zeta - X_0, \eta - W) d\zeta d\eta dX_0 dX_1 dW = 0$$

Let $\mu = \zeta - X_0$ and $\nu = \eta - W$. Then:

$$\int\int\int\int\int e^{it(\zeta-\mu)} e^{is(\eta-\nu)} e^{irX_1} \phi(\zeta, \eta, X_1) f_{U,V}(\mu, \nu) d\zeta d\eta dX_1 d\mu d\nu = 0$$

$$\underbrace{\int\int\int e^{(it\zeta+is\eta+irX_1)} \phi(\zeta, \eta) d\zeta d\eta dX_1}_{\mathcal{F}_\phi(t,s,r)} \underbrace{\int\int e^{-i(t\mu+s\nu)} f_{U,V}(\mu, \nu) d\mu d\nu}_{\bar{\mathcal{F}}_f(t,s)} = 0 \tag{23}$$

Equation 23 can be equal to zero if $\mathcal{F}_\phi(t, s, r)$ or $\bar{\mathcal{F}}_f(t, s)$ equals zero. By Assumption 18, $\bar{\mathcal{F}}_f(t, s)$ is different from zero, hence $\mathcal{F}_\phi(t, s, r) = 0$. Since the Fourier transform is injective, this implies that $\phi(\zeta, \eta, X_1) = 0$ and thus (Y, Z, X_1) is complete for (X, W). ■

Note that Assumption 18 is stronger than we require in our identification theorem, and we present Lemma 10 just as an illustration of the completeness assumption. Moreover, Assumption 18 can be relaxed under a location-scale model. Proof of completeness with such a construction can be found in Hu and Shiu (2011).

B Proofs of Theorems

B.1 Theorem 1

Proof. By Assumption 1

$$\mathbb{E}[H(Y) - \varphi(Z) - X|X, W] = 0$$

Let us recall two more functions $H^*(Y)$ and $\varphi^*(Z)$. By Assumption 1 again, we can write:

$$\mathbb{E}[H(Y) - \varphi(Z) - X|X, W] = 0 \quad \mathbb{E}[H^*(Y) - \varphi^*(Z) - X|X, W] = 0$$

If we take the difference of the two expectations:

$$\mathbb{E}[(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) + (X - X)|X, W] = 0$$

then by Assumption 2.1:

$$(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) = 0$$

by Assumption 3:

$$(H(Y) - H^*(Y)) = (\varphi(Z) - \varphi^*(Z)) = c$$

finally by Assumption 4:

$$c = 0$$

then:

$$H(Y) = H^*(Y) \quad and \quad \varphi(Z) = \varphi^*(Z)$$

■

B.2 Lemma 3

Proof. Let $x \in \mathcal{G}_0$, then we can write: $\langle K_0 x, y\rangle = \langle Kx, y\rangle$. We can equally write: $\langle K_0 x, y\rangle = \langle x, K_0^* y\rangle$ and $\langle Kx, y\rangle = \langle x, K^* y\rangle$. Moreover, for $x \in \mathcal{G}_0$ and $z \in \mathcal{G}$, $\langle x, z\rangle = \langle x, \mathbb{P}z\rangle$, then $\langle x, K^* y\rangle = \langle x, \mathbb{P}K^* y\rangle = \langle x, K_0^* y\rangle$. Then $\mathbb{P}K^* = K_0^*$ ■

B.3 Theorem 5

Proof. Remember that the solution of our problem was given by

$$\Phi = (\alpha I + T^*T)^{-1} T^* X$$

For the proof, we will decompose our equation into three parts as was done in Darolles et al. (2011) and look at the rates of convergence term by term.

$$\hat{\Phi}^\alpha - \Phi = \underbrace{(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^* X - (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi}_{I}$$

$$+ \underbrace{(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi - (\alpha I + T^*T)^{-1}T^*T\Phi}_{II}$$

$$+ \underbrace{(\alpha I + T^*T)^{-1}T^*T\Phi - \Phi}_{III}$$

The first term (I) is the estimation error for the right-hand side (X) of the equation, the second term (II) is the estimation error coming from the kernels and the third term (III) is the regularization bias coming from the regularization parameter α.

First examine the first term:

$$I = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^* X - (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi$$

$$I = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*(X - \hat{T}\Phi)$$

$$\|I\|^2 = \left\|(\alpha I + \hat{T}^*\hat{T})^{-1}\right\|^2 \left\|\hat{T}^*X - \hat{T}^*\hat{T}\Phi\right\|^2$$

where the first term is $O_p\left(1/\alpha^2\right)$ by Darolles et al. (2011) and the second term is $O_p\left(N^{-1} + h_N^{2s}\right)$ by Assumption 7.

Now, let us look at the second term II:

$$II = (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\Phi - (\alpha I + T^*T)^{-1}T^*T\Phi$$

$$= \left[\left[I - (\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}\right] - \left[I - (\alpha I + T^*T)^{-1}T^*T\right]\right]\Phi$$

$$= \left[\alpha(\alpha I + \hat{T}^*\hat{T})^{-1} - \alpha(\alpha I + T^*T)^{-1}\right]\Phi$$

$$= (\alpha I + \hat{T}^*\hat{T})^{-1}(\hat{T}^*\hat{T} - T^*T)\alpha(\alpha I + T^*T)^{-1}\Phi$$

$$\|II\|^2 = \left\|(\alpha I + \hat{T}^*\hat{T})^{-1}\right\|^2 \left\|(\hat{T}^*\hat{T} - T^*T)\right\|^2 \left\|\alpha(\alpha I + T^*T)^{-1}\Phi\right\|^2$$

The first term in (II) is $O_p(1/\alpha^2)$ by Darolles et al. (2011) while the second one is of order $O_p\left((Nh_N^{p+2})^{-1} + h_N^{2s}\right)$ as a result of relation $\left\|\hat{T}^*\hat{T} - T^*T\right\| = O_p\left(max\left\|\hat{T} - T\right\|, \left\|\hat{T}^* - T^*\right\|\right)$ by Assumption 6 and by Florens et al. (2012). Finally, the third is equal to $O(\alpha^{(\nu\wedge 2)})$ by Darolles et al. (2011).

The third term can be examined more straightforwardly:

$$III = (\alpha I + T^*T)^{-1}T^*T\Phi - \Phi$$

$$= \Phi^\alpha - \Phi$$

and $\|III\|^2 = \|\Phi^\alpha - \Phi\|^2$ is $O(\alpha^{\nu\wedge 2})$ by Assumption 5. Finally if we combine all terms we have:

$$\left\|\hat{\Phi}_N^\alpha - \Phi\right\|^2 = O_p\left(\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2s}\right) + \frac{1}{\alpha^2}\left(\frac{1}{Nh_N^{p+2}} + h_N^{2s}\right)\left(\alpha^{(\nu\wedge 2)}\right) + \alpha^{(\nu\wedge 2)}\right)$$

The proof of the second part of the theorem follows by Assumption 8. ■

B.4 Theorem 8

Proof.

$$H(Y) - \varphi(Z) - X_0 - X_1'\beta = U$$

$$\mathbb{E}[H(Y) - \varphi(Z) - X_0 - X_1'\beta|X, W] = 0 \quad by \quad Assumption \quad 1$$

Let us recall two more functions $H^*(Y)$, $\varphi^*(Z)$ and β^* such that:

$$H^*(Y) - \varphi^*(Z) - X_0 - X_1'\beta^* = U$$

Then, again by Assumption 1:

$$\mathbb{E}[H^*(Y) - \varphi^*(Z) - X_0 - X_1'\beta^*|X, W] = 0$$

If we take the difference of the two expectations:

$$\mathbb{E}[(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) - (X_1'\beta - X_1'\beta^*)|X, W] = 0$$

Then, by Assumption 9:

$$(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) - (X_1'\beta - X_1'\beta^*) = 0$$

By Assumption 10:

$$(H(Y) - H^*(Y)) - (\varphi(Z) - \varphi^*(Z)) = (X_1'\beta - X_1'\beta^*) = constant$$

$X_1'(\beta - \beta^*) = constant \to (\beta - \beta^*)'var(X_1)(\beta - \beta^*) = 0$. Then by Assumption 11, $\beta - \beta^* = 0$ implying $\beta = \beta^*$. Finally by Assumptions 3 and 4 we also get the identification of the functions of interest:

$$H(Y) = H^*(Y) \quad \varphi(Z) = \varphi^*(Z)$$

■

B.5 Theorem 9

Proof. Given the definition of β in equation (14), we use the following decomposition to prove the asymptotic normality of $\hat{\beta}$.

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \hat{M}_\alpha^{-1} \Bigg\{ \underbrace{\sqrt{N}[T_X^*(I - P_{YZ})\hat{E}(U|X, W)]}_{I} - \underbrace{\sqrt{N}[T_X^*(I - P_{YZ}) - \hat{T}_X^*(I - \hat{P}_{YZ}^\alpha)]\hat{E}(U|X, W)}_{II} \\ &\quad + \underbrace{\sqrt{N}[\hat{T}_X^*(I - \hat{P}_{YZ}^\alpha)\hat{T}(H, \varphi)]}_{III} \Bigg\}\end{aligned}$$

where $\hat{M}^\alpha = \hat{T}_X^*\hat{T}(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*\hat{T}_X - \hat{T}_X^*\hat{T}_X$, $P_{YZ} = T(T^*T)^{-1}T^*$, $\hat{P}_{YZ}^\alpha = \hat{T}(\alpha I + \hat{T}^*\hat{T})^{-1}\hat{T}^*$ and $\hat{E}(U|X, W) = X_0 - \hat{T}(H, \varphi) + \hat{T}_X\beta$. Given the assumptions we have introduced, in the sequel we prove the following:

- $\|\hat{M}_\alpha^{-1} - M^{-1}\| \to o_p(1)$ where $M = T_X^*T(T^*T)^{-1}T^*T_X - T_X^*T_X$
- $\|II\| \to O_p(1)$
- $\|III\| \to O_p(1)$

- $\hat{M}_\alpha^{-1}\left\{\sqrt{N}[T_X^*(I-P_{YZ})\hat{E}(U|X,W)]\right\} \to \mathcal{N}(0, \sigma^2 M^{-1}(\sum_{j/\psi_j \in \mathcal{R}(T)^\perp} E(X_1\psi_j)E(X_1\psi_j)')M^{-1})$

Proof of $\|\hat{M}_\alpha^{-1} - M^{-1}\|$

Note that we can equivalently write $\hat{M}_\alpha = \hat{T}_X^*(\hat{P}_{YZ}^\alpha - I)\hat{T}_X$ and $M = T_X^*(P_{YZ} - I)T_X$.

$$\|\hat{M}_\alpha^{-1} - M^{-1}\| \leq \|M^{-1}\|\|\hat{M}_\alpha^{-1}\|\|\hat{M}_\alpha - M\|$$

As M and $\hat{M}_\alpha$ are both finite rank operators, their inverses are bounded. So, we need to show that $\|\hat{M}_\alpha - M\| \to o_p(1)$.

$$\begin{aligned}
\|\hat{M}_\alpha - M\| &= \|\hat{T}_X^*(\hat{P}_{YZ}^\alpha - I)\hat{T}_X - T_X^*(P_{YZ} - I)T_X\| \\
&\leq \underbrace{\|\hat{T}_X^* - T_X^*\|\|(\hat{P}_{YZ}^\alpha - I)\hat{T}_X\|}_{A} + \underbrace{\|T_X^*[(\hat{P}_{YZ}^\alpha - I) - (P_{YZ}^\alpha - I)]\hat{T}_X\|}_{B} \\
&\quad + \underbrace{\|T_X^*(I - P_{YZ}^\alpha)\|\|\hat{T}_X - T_X\|}_{C} + \underbrace{\|T_X^*[(I - P_{YZ}^\alpha) - (I - P_{YZ})]T_X\|}_{D}
\end{aligned}$$

The second term in A is bounded in probability and the first term is of order $O_p(1/\sqrt{N})$. Extending results given in Florens et al. (2012) and using Assumptions 13 and 6.1, one can show that B is of order $O_p(\frac{1}{\sqrt{Nh^{p+q+1}}} + h^s)$. The second term in C is $o_p(1)$ which makes C $o_p(1)$. We can write D equivalently as $D = \|T_X^*[(I - P_{YZ}^\alpha)P_{YZ}]T_X\|$ which is of order $O(\alpha)$ by Florens et al. (2012) and Assumption 13. So, under the conditions given in the Assumption 8.1, $\|\hat{M}_\alpha^{-1} - M^{-1}\| \to o_p(1)$.

Proof of II:

$$\begin{aligned}
II &= \sqrt{N}[T_X^*(I - P_{YZ}) - \hat{T}_X^*(I - \hat{P}_{YZ}^\alpha)]\hat{E}(U|X,W) \\
&\leq \sqrt{N}[(T_X^* - \hat{T}_X^*)(I - \hat{P}_{YZ}^\alpha) - T_X^*(I - \hat{P}_{YZ}^\alpha) + T_X^*(I - P_{YZ})]\hat{E}(U|X,W) \\
\|II\| &\leq \underbrace{\sqrt{N}\|T_X^* - \hat{T}_X^*\|\|I - \hat{P}_{YZ}^\alpha\|\|\hat{E}(U|X,W)\|}_{A} + \underbrace{\sqrt{N}\|T_X^*(I - \hat{P}_{YZ}^\alpha) - T_X^*(I - P_{YZ})\|\hat{E}(U|X,W)\|}_{B}
\end{aligned}$$

A is $O_p(1)$. So, let us examine B in more detail.

$$\begin{aligned}
B &= \sqrt{N}\|T_X^*[(I - P_{YZ}) - (I - P_{YZ}^\alpha) + (I - P_{YZ}^\alpha) - (I - \hat{P}_{YZ}^\alpha)]\|\|\hat{E}(U|X,W)\| \\
&\leq \underbrace{\sqrt{N}\|T_X^*[(I - P_{YZ}) - (I - P_{YZ}^\alpha)]\|\hat{E}(U|X,W)\|}_{B1} \\
&\quad + \underbrace{\sqrt{N}\|T_X^*[(I - P_{YZ}^\alpha) - (I - \hat{P}_{YZ}^\alpha)]\|\|\hat{E}(U|X,W)\|}_{B2}
\end{aligned}$$

B1 can be written as $B1 = \sqrt{N}\|T_X^*(I-P_{YZ}^\alpha)P_{YZ}\hat{E}(U|X,W)\|$ and it is of order $O_p(\sqrt{N\alpha}(\frac{1}{\sqrt{Nh^{p+q}}}+h^s))$. We need to manipulate B2 a bit more to get its final rate of convergence.

$$\begin{aligned}
B2 &= \sqrt{N}\|T_X^*[(I-P_{YZ}^\alpha)-(I-\hat{P}_{YZ}^\alpha)]\|\|\hat{E}(U|X,W)\| \\
&\leq \sqrt{N}\|T_X^*(I-\hat{T}(\alpha I+\hat{T}^*\hat{T})^{-1}\hat{T}^* - T_X^*(I-T(\alpha I+T^*T)^{-1}T^*\|\|\hat{E}(U|X,W)\| \\
&\leq \sqrt{N}\|(I-(\alpha I+\hat{T}\hat{T}^*)^{-1}\hat{T}\hat{T}^*T_X-(I-(\alpha I+TT^*)^{-1}TT^*T_X\|\|\hat{E}(U|X,W)\| \\
&\leq \sqrt{N}\|\alpha[(\alpha I+\hat{T}\hat{T}^*)^{-1}-(\alpha I+TT^*)^{-1}]T_X\|\|\hat{E}(U|X,W)\| \\
&\leq \sqrt{N}\|\alpha[(\alpha I+\hat{T}\hat{T}^*)^{-1}(TT^*-\hat{T}\hat{T}^*)]T_X\|\|\hat{E}(U|X,W)\| \\
&\leq \sqrt{N}\|\alpha(\alpha I+\hat{T}\hat{T}^*)^{-1}\hat{T}(\hat{T}^*-T^*)T_X\hat{E}(U|X,W)\| \\
&\quad +\sqrt{N}\|\alpha(\alpha I+\hat{T}\hat{T}^*)^{-1}(\hat{T}-T)T^*T_X\hat{E}(U|X,W)\|
\end{aligned}$$

The first term is of order $O_p(\sqrt{\alpha N}(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s))$ by Darolles et al. (2011) and Assumption 6.1 and the second term is of order $O_p(\sqrt{N}\alpha(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s))$ again by Darolles et al. (2011) and Assumption 6.1. Thus II is $O_p(1)$ under the conditions given in the Assumption 14.

Proof of III:

$$III = \sqrt{N}[\hat{T}_X^*(I-\hat{P}_{YZ}^\alpha)\hat{T}(H,\varphi)]$$

$$\begin{aligned}
\|III\| &= \sqrt{N}\|\hat{T}_X^*(I-\hat{P}_{YZ}^\alpha)\hat{T}(H,\varphi)\| \\
&= \sqrt{N}\|\hat{T}_X^*\|\|(I-\hat{P}_{YZ}^\alpha)\hat{T}\|\|(T^*T)^{\nu/2}-(\hat{T}^*\hat{T})^{\nu/2}\|\|g\| \\
&\quad +\sqrt{N}\|\hat{T}_X^*\|\|(I-\hat{P}_{YZ}^\alpha)\hat{T}(\hat{T}^*\hat{T})^{\nu/2}\|\|g\|
\end{aligned}$$

where $g \in \mathcal{E}_0$ such that $(H,\varphi) = (\hat{T}^*\hat{T})^{\nu/2}g$ for some $\nu > 0$. $\|(I-\hat{P}_{YZ}^\alpha)\hat{T}\|$ is of order $O_p(\sqrt{\alpha})$ by Florens et al. (2012). Note that following Engl et al. (1996), we can write: $\|(T^*T)^{\nu/2}-(\hat{T}^*\hat{T})^{\nu/2}\| \leq \|T^*T-\hat{T}^*\hat{T}\|^{\min\{\nu,2\}/2}$. Then it is of order $O_p(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s)$ by Assumptions 13 and 6.1. The second line is of order $O_p(\sqrt{N\alpha})$ by Florens et al. (2012) and Engl et al. (1996). So the third term is of order $O_p(\sqrt{N\alpha}(\frac{1}{\sqrt{Nh^{p+q+1}}}+h^s)+\sqrt{N\alpha})$. Given the conditions in Assumption 14, III is of order $O_p(1)$.

Proof of $\hat{M}_\alpha^{-1}\left\{\sqrt{N}[T_X^*(I-P_{YZ})\hat{E}(U|X,W)]\right\}$

We have already shown that $\|\hat{M}_\alpha^{-1}-M^{-1}\| = o_p(1)$. Let us now look at term I:

$$I = \sqrt{N}\hat{T}_X^*(I-P_{YZ})\hat{E}(U|X,W)$$

$$I = \frac{\sqrt{N}}{N}\sum_i u_i \sum_{j/\psi_j\in\mathcal{R}(T)^\perp}\langle\frac{K_N(x-x_i)K_N(w-w_i)}{\hat{f}(x,w)},\psi_j\rangle E(X_1\psi_j)$$

This decomposition is based on the following property: $\forall\psi\in L_F^2(X,W)$, $\psi = \sum_j\langle\psi,\psi_j\rangle\psi_j$, then $(I-P_{YZ})\psi = \sum_{j/\psi_j\in\mathcal{R}(T)^\perp}\langle\psi,\psi_j\rangle\psi_j$. Moreover, using a standard argument in Kernel estimation,

we may replace $\hat{f}(x,w)$ in the denominator by the true $f(x,w)$:

$$I = \frac{\sqrt{N}}{N}\sum_i u_i \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)}, \psi_j \rangle E(X_1\psi_j)$$

Let us denote

$$\mu_i = u_i \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)}, \psi_j \rangle E(X_1\psi_j)$$

If μ_i is a random variable with mean zero and a finite variance we can apply the CLT to obtain the asymptotic normality for $\hat{\beta}$. However, note that μ_i also depends on the sample size N. For this reason, besides showing that it has a finite mean and variance, we also need to show that it satisfies the Liapounoff condition to apply the Liapounoff central limit theorem, see Pollard (1984); Amemiya (1986). It is straightforward to show that $E(\mu_{iN}) = 0$.

$$Var(\mu_{iN}) = E(Var(\mu_{iN}|X,W)) + Var(E(\mu_{iN}|X,W)) = E(Var(\mu_{iN}|X,W))$$

$$\begin{aligned} Var(\mu_{iN}) &= \sigma^2 E\left\{ \left[\sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)}, \psi_j \rangle E(X_1\psi_j) \right] \right. \\ &\quad \left. \times \left[\sum_{l/\psi_l \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)}, \psi_l \rangle E(X_1\psi_l) \right]' \right\} \\ &= \sigma^2 \sum_{j,l} E\{ \langle \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)}, \psi_j \rangle \langle \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)}, \psi_l \rangle \} E(X_1\psi_j)E(X_1\psi_l)' \end{aligned}$$

Let us write the first expectation for $j \neq l$:

$$\begin{aligned} &= \int f(x_i,w_i)dx_i dw_i \int \frac{K_N(x-x_i)K_N(w-w_i)}{f(x,w)} \psi_j(x,w) f(x,w) dx dw \\ &\quad \times \int \frac{K_N(\tilde{x}-x_i)K_N(\tilde{w}-w_i)}{f(\tilde{x},\tilde{w})} \psi_l(\tilde{x},\tilde{w}) f(\tilde{x},\tilde{w}) d\tilde{x} d\tilde{w} \\ &= \int f(x_i,w_i)\psi_j(x_i,w_i)\psi_l(x_i,w_i) \\ &= 0 \end{aligned}$$

since the ψ_j's are orthogonal. Hence the covariance term is zero. Now let us check the variance, i.e., the case where $j = l$:

$$
\begin{aligned}
&= \int f(x_i, w_i) dx_i dw_i \left\{ \int \frac{K_N(x - x_i) K_N(w - w_i)}{f(x, w)} \psi_j(x, w) f(x, w) dx dw \right\}^2 \\
&= \int \psi_j^2(x_i, w_i) f(x_i, w_i) dx_i dw_i \\
&= E(\psi_j^2(x_i, w_i)) = 1
\end{aligned}
$$

which is equal to 1 since ψ_j is an orthonormal base. Then we can conclude that:

$$
Var(\mu_{iN}) = \sigma^2 \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} E(X_1 \psi_j) E(X_1 \psi_j)'
$$

We now have to check the Liapounoff condition (See Pollard, 1984, Chapter 3):
Let $X_{i1}, X_{i2}, ..., X_{iN}$, be independent random variables with zero means and variances that sum to one. Then the Liapounoff condition is $\lim_{N \to \infty} \sum_{i=1}^{N} E[|X_{iN}|^{2+\delta}] = 0$, for $\delta > 0$.
Let us denote:

$$
X_{iN} = \frac{\frac{\sqrt{N}}{N} \sum_i u_i \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x - x_i) K_N(w - w_i)}{f(x,w)}, \psi_j \rangle E(X_1 \psi_j)}{\sqrt{Var(\sum_i u_i \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x - x_i) K_N(w - w_i)}{f(x,w)}, \psi_j \rangle E(X_1 \psi_j))}}
$$

We need to show that X_{iN} satisfies the Liapounoff condition. Without loss of generality, let us assume that $\delta = 1$. Then we can write:

$$
lim_{N \to \infty} E \left[\frac{\left| \frac{\sqrt{N}}{N} \sum_i u_i \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x - x_i) K_N(w - w_i)}{f(x,w)}, \psi_j \rangle E(X_1 \psi_j) \right|^3}{\left| Var(\sum_i u_i \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \langle \frac{K_N(x - x_i) K_N(w - w_i)}{f(x,w)}, \psi_j \rangle E(X_1 \psi_j)) \right|^{3/2}} \right] = 0
$$

We have already shown that the variance exists and is finite. So let us examine the numerator. By Assumption 16, we can write:

$$
\begin{aligned}
&= \frac{1}{\sqrt{N}} E[|u_i|^3 | X, W] \frac{1}{N} \sum_{i=1}^{N} E \left[\left| \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \int \frac{K_N(x - x_i) K_N(w - w_i)}{f(x, w)} \psi_j(x, w) f(x, w) dx dw E(X_1 \psi_j) \right|^3 \right] \\
&= \frac{1}{\sqrt{N}} E[|u_i|^3 | X, W] \frac{1}{N} \sum_{i=1}^{N} E \left[\left| \sum_{j/\psi_j \in \mathcal{R}(T)^\perp} E(X_1 \psi_j) \psi_j \right|^3 \right] \\
&= \frac{1}{\sqrt{N}} E[|u_i|^3 | X, W] \frac{1}{N} \sum_{i=1}^{N} E \left[|(I - P_{YZ}) X_1|^3 \right]
\end{aligned}
$$

By Assumption 16, we can then conclude that $\lim_{N\to\inf}\sum_{i=1}^{N} E[|X_{iN}|^{2+\delta}] = 0$ and the Liapounoff condition is satisfied. Then by applying the Liapounoff central limit theorem, we can write:

$$\sqrt{N}\hat{M}_{\alpha}^{-1}T_X(I - P_{YZ})\hat{E}(U|X,W) \to \mathcal{N}(0, \sigma^2 M^{-1} \sum_{j/\psi_j \in \mathcal{R}(T)^{\perp}} E(X_1\psi_j)E(X_1\psi_j)'M^{-1})$$

We obtain the asymptotic normality of β by combining this result with the previous ones:

$$\sqrt{N}(\hat{\beta} - \beta) \to \mathcal{N}(0, \sigma^2 M^{-1} \sum_{j/\psi_j \in \mathcal{R}(T)^{\perp}} E(X_1\psi_j)E(X_1\psi_j)'M^{-1})$$

■

C Simulation Results

Table 1: Mean and Standard Error (in parenthesis) of the estimator of β

	Sample size		
	$N = 200$	$N = 500$	$N = 1000$
Nonparametric IV	0.2685	0.2756	0.2808
	(0.0371)	(0.0231)	(0.0169)

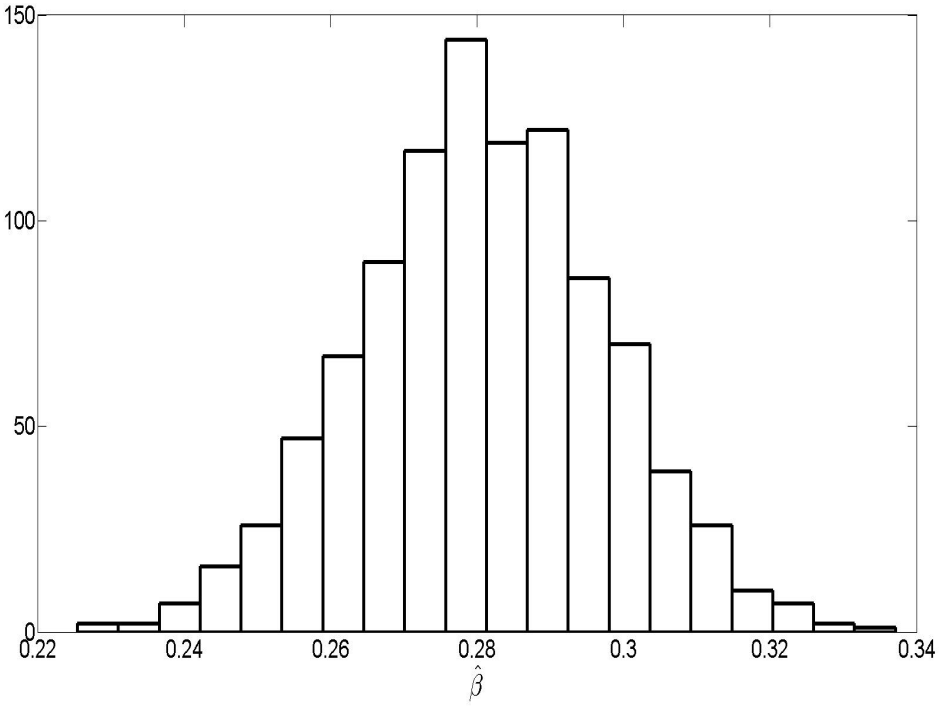

Figure C1: *Histogram of* $\hat{\beta}$. Histogram obtained from 1000 Monte Carlo replications for a sample size of 1000.

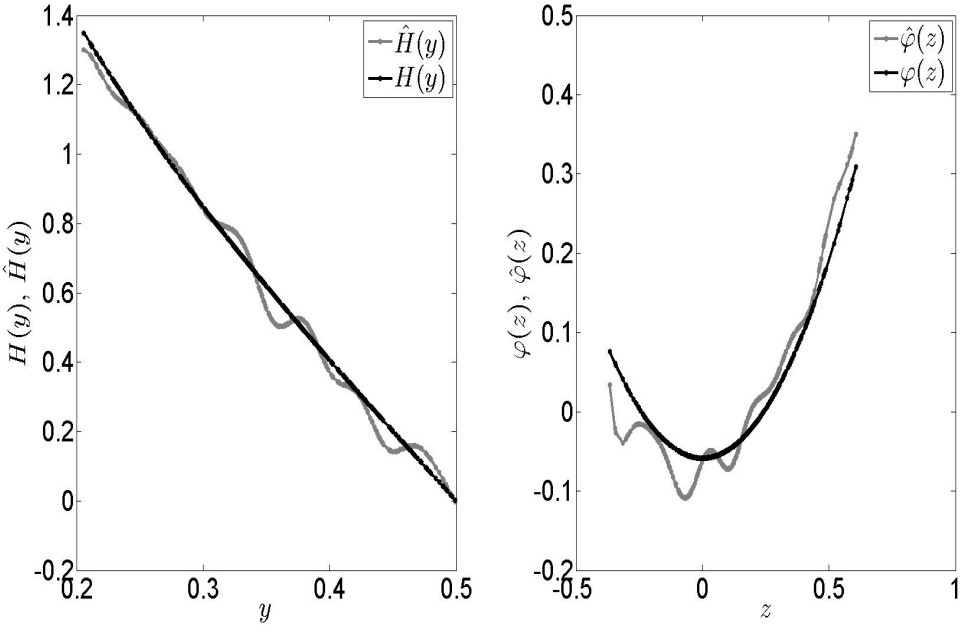

Figure C2: *Estimated functions for a sample size of* 500. Grey curves are the estimated functions whereas the black ones are the true functions. α_H and α_φ are chosen by the data driven rule given in *Section 4*.

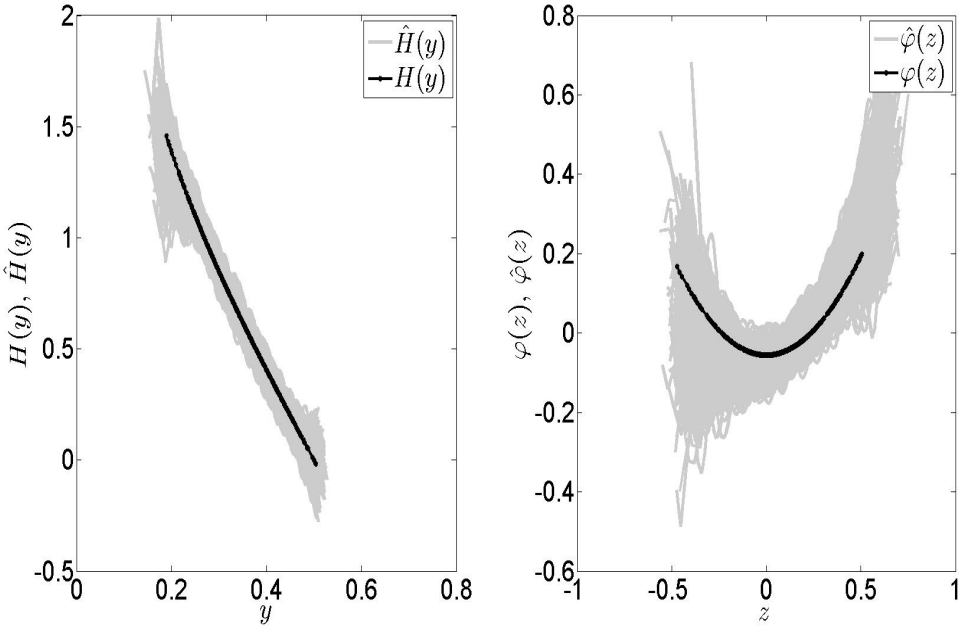

Figure C3: *Monte Carlo simulation.* Black curves are the true functions. Grey curves show the estimated functions at each simulation. α_H and α_φ are chosen by the data driven rule at each simulation.

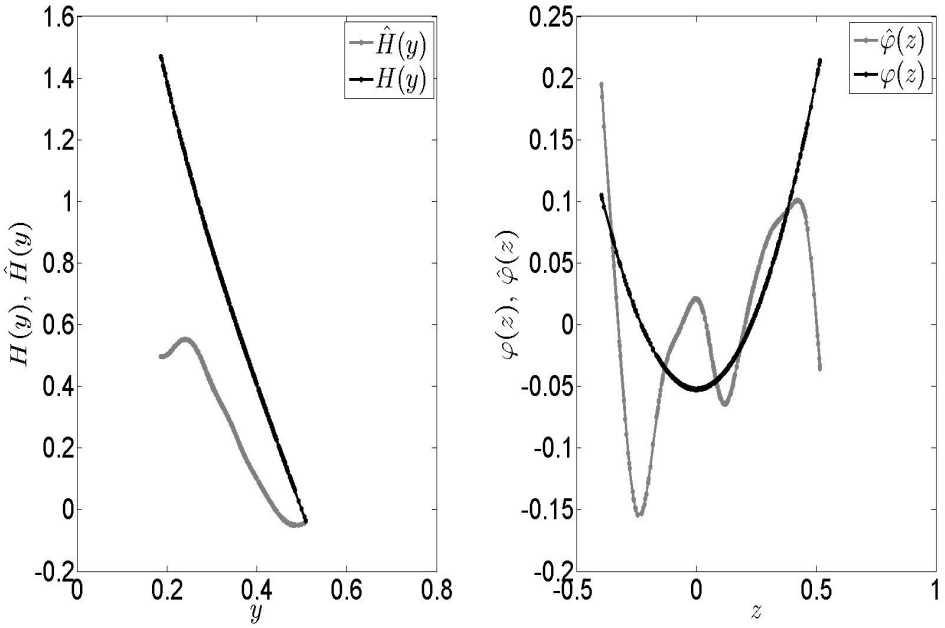

Figure C4: *Estimated functions for arbitrary α_H and α_φ.* Grey curves are the estimated functions whereas the black ones are the true functions.